Forecasting from time series subject to sporadic perturbations: Effectiveness of different types of forecasting support

Abstract

How effective are different approaches to providing forecasting support? Forecasts may be unaided or made with the help of statistical forecasts. In practice, the latter often comprise crude forecasts that do not take sporadic perturbations into account. In research, forecasts are generally based on series cleansed of perturbation effects. In an experiment, people made forecasts from time series disturbed by promotions. In all conditions, under-forecasting occurred on promotional periods and over-forecasting occurred on normal ones. Relative size of these effects depended on the proportion of periods containing promotions in the data series. Statistical forecasts improved forecasting accuracy not because they reduced these biases but because they decreased random error (scatter). Performance improvement did not depend on whether forecasts were based on cleansed series. Thus, efforts invested in producing forecasts based on cleaned time series may not be warranted: companies may benefit from giving their forecasters even crude statistical forecasts. In a second experiment, forecasters received optimal statistical forecasts that took effects of promotions fully into account. Accuracy was higher than before because biases were almost eliminated and random error was reduced by 20%. Thus the additional effort required to produce forecasts that take promotional effects into account is worthwhile.

Key words: Forecasting support; judgmental adjustment; time series; promotions; sales

1. Introduction

Business forecasters use both unaided judgmental forecasting and forecasting aided by formal statistical forecasts (Sanders & Manrodt, 2003). The latter approach may be increasing as users become more familiar with software that provides forecasting support. As a result, forecast support systems have great potential for improving forecast performance. However, there are factors that prevent this potential being fully realised. Forecasters tend to ignore the 'advice' provided by a formal forecast or take too little account of it (Goodwin, Fildes, Lawrence, & Nikolopoulos, 2007; Lim & O'Connor, 1996; Önkal, Goodwin, Thomson, Gönul, & Pollock, 2009). When they do take account of this advice, they do not assign enough weight to it. Consequently, the improvement in accuracy produced by it is generally small, albeit somewhat greater when series are complex and when the formal forecasts are of higher quality (Goodwin & Fildes, 1999; Goodwin, Fildes, Lawrence, & Stephens, 2011; Lim & O'Connor, 1995; Trapero, Pedregal, Fildes, & Kourentzes, 2013).

The picture is more complex in the case of series with sporadic perturbations, such as those associated with promotions. Goodwin and Fildes (1999) showed that, in this situation, statistical forecasts tend to help on normal periods but not on those subject to promotions. However, the statistical forecasts used in this research did not take effects of promotions into account: they were based on the baseline time series cleansed of the effects of promotions. Recently, forecasting models that do allow for the effects of promotions have been developed (Huang, Fildes, & Soopramanien, 2014; Kourentzes & Petropoulos, 2016; Trapero, et al., 2013). However, given that there is considerable lag between development of more sophisticated statistical models and their implementation by practitioners (Lawrence, 2000; Sanders & Manrodt, 2003), it is likely to be some time before they impact business practice.

Even in the case of relatively simple models, there appears to be a gap between the formal forecasts used in experimental studies and those used in business practice. In experimental studies, formal forecasts are based on non-promotional periods only (e.g., Goodwin & Fildes, 1999). In other words, they are calculated from the baseline series cleansed of promotion effects. In non-experimental studies, on the other hand, formal forecasts take no account of whether past periods contain promotions (Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009; Trapero, et al., 2013). Hence, if we are interested in the relevance of experimental results to business practice, we need to ask whether any advantage of using judgmentally adjusted statistical forecasts over unaided judgment depends on the type of statistical forecast used.

Goodwin and Fildes (1999) have argued that the benefit of providing statistical forecasts should be greater when they are based on data that have been cleansed of promotional effects. Referring to the estimated level of sales when a promotion does not run as the *baseline* value, they point out that this is because the baseline values provided by that type of statistical forecast can be accepted without any adjustment when no promotions are planned. Moreover, past differences between promotional and non-promotional periods can be directly used as a basis for assessing the size of the adjustment needed when promotions are planned.

In what follows, we address the following questions. First, is there an advantage of using a judgmentally adjusted statistical forecast over using unaided judgment? Second, is any such advantage greater when statistical forecasts are based on past data cleansed of promotional effects? Third, does any benefit derived from provision of statistical forecasts depend on features of the data series (i.e., ratio of promotional to non-promotional periods) or of the periods to be forecast (i.e., whether a promotion is planned)? Finally, can people make good use of ‘ideal’ statistical forecasts that include allowance for the effects of promotions (cf., Huang et al, 2014; Kourentzes and Petropoulos, 2016; Trapero et al, 2013)? In other words, if their goal is to maximize forecasting accuracy, do they adopt these forecasts without making any adjustment?

2. Development of hypotheses

In their survey, Fildes and Goodwin (2007) found that 75% of respondents indicated that they used judgment when making forecasts. Of these, 25% said that they used unaided judgment and 50% said that they used a combination of judgment and statistical forecasting (averaging, judgmental adjustment). Over recent years, use of statistical software has become more pervasive in business settings and so the proportion of forecasters using a combinatorial approach has increased further: it had risen to 55% by 2014 (Fildes & Petropoulos, 2015).

Judgmental adjustment does not always improve statistical forecasts. People tend to make *unnecessary* adjustments even when they have no additional information (Goodwin, 2000; Lawrence, Goodwin, O'Connor, & Önkal, 2006). This may be because they discern patterns in noise (Fildes, et al., 2009), because they are too optimistic and place excess weight on positive signals (Bovi, 2009; Durand, 2003; Kotteman, Davis, & Remus, 1994), or because they want to feel ownership of their forecasts (Önkal & Gönul, 2005). They also tend to be overconfident in the accuracy of their forecasts (Arkes, 2001; Bovi, 2009; Lawrence, et al., 2006), perhaps because a self-serving attribution bias causes them to overestimate the importance of their own judgment relative to that of the statistical forecast (Hilary & Hsu, 2011; Libby & Rennekamp, 2012).

All these studies have focused on whether judgmentally adjusted forecasts are better or worse than raw statistical forecasts. The underlying issue was whether forecasters should be allowed to make adjustments to statistical forecasts and, if they should, whether anything can be done to ensure that their adjustments are beneficial (Goodwin, et al., 2011). In contrast, our primary aim here is to investigate the value of providing a formal forecast to increase forecasting accuracy. Thus, our main focus is on whether judgmentally adjusted statistical forecasts are better or worse than unaided judgmental forecasts[1]. For us, the underlying issue is to quantify the benefit of providing forecasters with forecasting support (operationalized in this paper as the provision of a statistical forecast, including historic forecasts). Such support has been assumed to be beneficial (Alvarado-Valencia & Barrero, 2014) because it reduces the processing demands imposed on forecasters (Fildes & Goodwin, 2013). Furthermore, combining forecasts from more than one source outperforms the results of a single forecasting method (Armstrong, 2001), particularly when the two methods are independent and rely on different information. Given the complementary nature of judgment and statistical methods, their combination should be especially beneficial (Blattberg & Hoch, 1990). Therefore:

H1: Providing forecasters with statistical forecasts improves forecasting accuracy compared to unaided judgment.

Önkal, Sayim, & Lawrence (2012) noted that some differences exist between the characteristics of forecasts examined in experimental research and those prevalent in business practice. As mentioned above, one such difference is in the nature of the statistical forecast provided when series are subject to perturbations of the sort typically produced by promotions: in experimental work, statistical forecasts have been based on series cleansed of promotional effects (Goodwin and Fildes, 1999; Goodwin et al, 2011) whereas, in business data analysed by researchers, they have not (Fildes, et al., 2009; Trapero, et al., 2013). As we mentioned above, Goodwin and Fildes (1999) expected the former approach to produce better results. Specifically, they argued: "This has the benefit of clearly separating the underlying time series from the promotion effects. Moreover, some commercial forecasting packages like *Forecast Pro* now allow observations for special periods to be separated out so that they cannot contaminate forecasts for normal periods. … With access to a statistical time series forecast of the 'baseline value' the judge has only to estimate the effect of the cue and make an appropriate adjustment to the statistical forecast " (p 41).

Insert Figure 1 about here

As an example, consider a promotion of a given size that has elevated sales by 100 units above the baseline in the past. If a promotion of the same size is planned for the future, 100 units can simply be added to the statistical forecast of the baseline forecast (Figure 1, Lower panel). On the other hand, if no promotion is planned, the baseline forecast can be adopted without adjustment. In contrast, statistical forecasts based on non-cleansed data always have to be adjusted. When no promotion is planned, the forecast must be adjusted downwards and, when one is planned, it must be adjusted upwards (Figure 1, Upper panel). Forecasters need to know how much the statistical forecast has been influenced by the presence of past promotions in the data series. Without that knowledge, it is difficult for them to know how much to adjust upwards when promotions are planned and how much to adjust downwards when they are not. Thus, the forecasting process is more complex than when the statistical forecast is based on cleansed data series.

In fact, few studies have compared the effects of different types of statistical forecast on the accuracy of judgmental forecasters provided with those forecasts. One exception is Lim and O'Connor's (1995) experimental study of forecasting from time series without disturbances. They manipulated the accuracy of the statistical forecasts; this varied from low (naïve forecast) to medium (forecast produced by damped exponential smoothing) to high (average of the actual value and the forecast produced by damped exponential smoothing). Participants were asked to make an initial forecast based on their own judgment and were then presented with one of the three types of statistical forecast. After every trial, they were able to see their final forecast, the statistical forecast and the actual value, thus facilitating learning over trials. There was an overall beneficial effect of providing statistical forecasts, consistent with our first hypothesis. Additionally, more accurate statistical forecasts provided greater improvements in accuracy.

Thus, based on Goodwin and Fildes (1999) reasoning and on Lim and O'Connor's (1995) findings:

Hypothesis 2: Formal forecasts based on cleansed series are more beneficial than those based on non-cleansed series.

Judgmental forecasting from time series appears to depend on use of anchoring heuristics (Lawrence & O'Connor, 1992). Given an un-trended data series that includes both normal and promotional periods, unaided forecasters are likely to anchor on the mean of that series. Then they adjust upwards to allow for the presence of a planned promotion in the forecast period and adjust downwards to allow for the absence of a planned promotion (Figure 1, Upper panel). Given that adjustment is typically insufficient when anchoring heuristics are used (Tversky & Kahneman, 1974), we expect under-forecasting for promotional periods but over-forecasting for normal ones. As statistical forecasts based on non-cleansed series follow the mean of the data series, we expect the same mental anchor to be used as for unaided forecasting. Thus, where directional error is given by the outcome minus the forecast:

H3a*: For forecasting that is unaided or aided by statistical forecasts based on non-cleansed data series, directional error will be positive for normal periods and negative for promotional ones.*

However, when statistical forecasts are based on cleansed data series, the mean of the statistical forecast history will approximate the mean of the non-promotional periods. Hence, to predict sales for a period when no promotion is planned, forecasters do not need any adjustment. However, for promotional periods, they still need to adjust upwards (Figure 1, Lower panel) and this adjustment will be insufficient. Hence:

H3b: For forecasting that is aided by statistical forecasts based on cleansed data series, the directional error will be zero for normal periods and negative for promotional periods.

Statistical forecasts based on non-cleansed series tend to lie between the sales level associated with non-promotional periods and the average sales level associated with promotional periods. When the ratio of promotional to non-promotional periods is low (e.g., 10%), the historical mean of statistical forecasts will be much closer to the actual baseline of the series than when it is high (e.g., 40%). This should benefit forecasting for periods without promotions as minimal adjustment is required. However, when this ratio is low, there is less information on which to estimate the relation between promotional size and its effect. This is likely to impair forecasting for promotional periods. Thus, when statistical forecasts are based on non-cleansed series:

H4: A lower proportion of promotions in the data series will benefit forecasts for non-promotional periods but impair those for promotional periods.

When there are relatively few promotional periods in the data, statistical forecasts based on non-cleansed data series are closer to the baseline and, as a result, they approximate statistical forecasts based on cleansed data. In contrast, when the proportion of promotional periods is high, statistical forecasts based on non-cleansed series are well above the baseline and the difference between them and statistical forecasts based on cleansed-series is larger. Hence:

Hypothesis 5: Any difference in the benefits derived from the two types of statistical forecast will be greater when the proportion of promotional periods in the data series is higher.

3. Experiment 1: Use of statistical forecasts that take no account of effects of promotions

A mixed design was used to test these hypotheses. Type of task (unaided judgmental forecasting/forecasting aided by statistical forecasts based on non-cleansed series/forecasting aided by statistical forecasts based on cleansed series) was varied between participants and proportion of promotions in the presented data (40% versus 10%) and forecasting for promotional versus non-promotional periods were varied within participants.

3.1. Method

3.1.1. Participants A total of 153 students from University College London participated in the study. Their mean age was 18.56 years (*SD* = 1.03 years) and 127 of them were female.

3.1.2 Design and stimulus materials Forty series, each consisting of 50 data points, were generated with R statistical software. Half of the series were independent (mean = 300, error = 7%) and half were autoregressive (mean = 300, ρ = 0.7, error = 7%). Series were displayed as a grey line and were labelled ‘sales’. The graphs also contained vertical blue bars[2] that indicated promotional expenditure on either five or 20 of the 50 periods. Both location and size of these promotions were assigned randomly. Promotion size was selected at random without replacement from a list of every tenth value between 50 and 200. The size of the promotion had a same-week effect on the sales number according to the following formula:

$$PI_t = \frac{Pt}{5} * S_t$$

This indicates a same-week percentage increase *PI* at time *t* (over the regular sales *S* at that time) equal to one fifth of the promotional expenditure *P*.

Participants were asked to forecast one step ahead (period 51) and two steps ahead (period 52). A promotion was present either on time period 51 or time period 52. The size of this promotion was randomized for every participant across trials. Over the experimental session, it included every tenth value between 30 and 220. Thus, participants were required to forecast four promotion sizes (30, 40, 210, 220) that were not included in the range presented in the data series (i.e., 50–200).

Insert Figure 2 about here

The presence and type of statistical forecasts was manipulated between participants. The first group (A) did not receive a statistical forecast (Figure 2a). Two other groups received a statistical forecast calculated using the Holt-Winters exponential smoothing method. A line graph represented the statistical forecast history for week 2 to week 52. One of these groups (B) received a statistical forecast based on the baseline data series, cleansed of promotional effects (Figure 2b). The other group (C) received a statistical forecast based on the total sales: in calculating it, no distinction was made between normal and promotional periods (Figure 2c).

3.1.3. Procedure Participants were given instruction sheets, which differed only on the explanation of the statistical forecast (see Appendix). In addition, participants were orally instructed to pay close attention to the explanation of the graphical components on their instruction sheet and were given a short demonstration of two example trials.

3.2. Results

We present analyses of three error scores[3]: mean absolute error (MAE), mean error (ME), and variable error (VE).

3.2.1. Mean absolute error MAE was used to measure overall error level. Errors were calculated relative to the ideal forecast. This was provided by the *signal* (excluding the noise) of the time series on non-promotional periods and by the signal plus the promotion effect on promotional periods. For the independent time series, the ideal forecast for a non-promotional period was 300 (the mean). For the autoregressive series, it was 0.2 (1-ρ) of the distance from the last data point towards the mean. Outlier analysis indicated two participants had scored more than two standard deviations away from the mean on half of the trials or more; they were therefore excluded from the analyses.

Table 1 shows MAE values for each combination of the three independent variables. The overall mean value of MAE was 30.20 (*SD* = 9.65).

Insert Table 1 about here

An analysis of variance with statistical forecast as a between-participants variable and promotion frequency and promotion presence as within-participant variables revealed a main effect of statistical forecast (F (2,150) = 3.99, p = .021, η_p^2 = .050). Hypothesis 1 stated that the provision of a statistical forecast would be beneficial to forecasting accuracy, such that unaided judgment would result in higher error than aided judgment. One-tailed t-tests confirm that the MAE for the unaided judgment group (*MAE* = 33.26, *SD* = 9.84) was significantly higher from that of the cleansed forecast group (*MAE* = 28.59, *SD* = 9.29; t (100) = 2.47, p = .008) and significantly higher than that of the non-cleansed forecast (*MAE* = 28.77, *SD* = 9.26; t (100) = 2.37, p = .010).

The MAE scores of the cleansed forecast group and the non-cleansed forecast group were not significantly different from one another (t (100) = -.10, p = .921). Thus we failed to obtain support for Hypothesis 2, which stated that participants given a statistical forecast based on cleansed series would be more accurate than those given a statistical forecast based on non-cleansed series.

There was a main effect of frequency of promotions in the data series (F (2,150) = 28.21, p < .001, η_p^2 = .158), a main effect of the presence of a promotion in the period to be forecast (F (2,150) = 7.35, p = .008, η_p^2 = .047), and an interaction between these two variables (F (2,150) = 743.82, p < .001, η_p^2 = .226). Analysis of simple effects showed that these effects arose because lower error with less frequent promotions occurred when forecasts were made for non-promotional periods but not when they were made for promotional ones (*F* (1, 150) = 61.66, *p* < .001).

We failed to obtain support for Hypothesis 5: there was no significant interaction between the type of statistical forecast provided and frequency of promotions in the data series.

3.2.2. Mean error The MAE score discussed above is a measure of overall error. Following Thurstone (1926), overall error can be decomposed into directional error or bias (ME) and scatter or variable error (VE). Taking D as the Forecast - Actual, ME is defined as ΣD/n and VE as √ ([Σ (D – ME)^2]/n). Thus, overall error could theoretically comprise a) bias but no scatter (all forecasts are a fixed distance from the optimal forecast with no distribution around that point), b) scatter but no bias (forecasts are distributed around a central point but that central point is the optimal forecast), or c) bias and scatter (forecasts are distributed around a central point that is a fixed distance from the optimal forecast). In practice, both bias and scatter contribute to overall error but their relative contributions depend on contextual factors.

To investigate the reasons for the differences in MAE reported above and to test hypothesis 3 – 5, we report analyses of ME in this section and of VE in the following one (Table 2).

Insert Table 2 about here

No effects involving the statistical forecast variable were significant. There was a main effect of whether the forecast was for a period with promotions (*F* (2,150) = 126.69, *p* < .001, η_p^2 = .458). Mean Error was negative when forecasts were for periods on which promotions were planned but positive when they were for periods with no promotions planned. There was also a main effect of the proportion of promotions in the data series (*F* (2,150) = 67.17, *p* < .001, η_p^2 = .309): overall, ME was lower when there was a low proportion of promotions in the data series than when there was a high one. There was also a significant interaction between these two variables (*F* (2,150) = 6.49, *p* = .012, η_p^2 = .041). Analysis of simple effects showed that this arose because a lower proportion of promotions in the data series decreased the positive ME of forecasts for non-promotional periods (*F* (1, 150) = 63.77, $p < .001$) but increased the negative ME of forecasts for promotional periods (*F* (1, 150) = 16.54, $p < .001$). This result is consistent with Hypothesis 4 that stated that fewer promotions would benefit forecasts for non-promotional periods but impair those for promotional ones.

Hypothesis 3a predicted that, when statistical forecasts were based on *non-cleansed* series, we would observe under-forecasting for promotional periods but over-forecasting for normal periods. One-sample t-tests confirmed that ME was significantly below zero on promotional periods (*t* (50) = -2.27, *p* = .028) and significantly above zero (*t* (50) = 8.20, $p < .001$) on normal ones.

Hypothesis 3b predicted that, for the forecasts based on *cleansed* series, the ME for normal periods would be zero and the ME for promotional periods would be negative (i.e., under-forecasting). While the ME for promotional periods in the cleansed series condition was indeed significantly below zero (*t* (50) = -3.51, *p* = .001), the ME for normal periods was positive and significantly different from zero (*t* (50) = 7.37, $p < .001$). This unexpected over-forecasting on normal periods was greater when there were 40% promotions in the data series than when there were 10% promotion in the data series (*t* (50) = 4.25, $p < .001$).

3.2.3. Variable Error. There was a significant effect of group on VE (*F* (2,150) = 3.56, *p* = .031). We hypothesized that the error of the unaided judgment group would be higher than that of the aided judgment groups. One tailed t-tests confirm that the VE of unaided judgment group was indeed larger than that of the group who received cleansed forecasts (*t* (100) = 2.04, *p* = .044), and that of the group that received non cleansed forecasts (*t* (100) = 2.53, *p* = .013). (VE scores in the latter two groups were not significantly different from one another.)

Forecasts from data series with 40% promotions had higher VE scores than those from data series with 10% promotions (*F* (1,150) = 10.23, *p* = .002, $\eta_p^2 = .064$). In addition, forecasts for promotional periods had higher VE than those for non-promotional ones (*F* (1,150) = 24.66, $p < .001$, $\eta_p^2 = .141$). Additionally, there was an interaction effect between these two variables (*F* (1,150) = 4.84, *p* = .029, $\eta_p^2 = .031$). Analysis of simple effects indicated that the error difference between normal and promotional periods was more pronounced in the low promotion frequency trials (*F* (1,150) = 29.26, $p < .001$, $\eta_p^2 = .163$) than in the high promotion frequency trials (*F* (1,150) = 7.37, $p = .007$, $\eta_p^2 = .047$).

3.3. Discussion

The experiment produced two separate groups of effects. The first concerns the effects on MAE and VE of whether participants made unaided forecasts, made forecasts after being given non-cleansed statistical forecasts, or made forecasts after being given cleansed statistical forecasts. The second concerns effects on MAE, ME, and VE of the proportion of promotional periods in the data series and of whether forecasts were made for promotional or normal periods. As there were no interactions between these two groups of effects, we will discuss them separately. Once we have done so, we will summarise a unitary account of the cognitive processes underlying performance that explains both types of effect.

3.3.1. Effects of providing forecast support Provision of statistical forecasts reduced overall error (MAE). However, further analysis showed that this was not because they reduced the directional error or bias (ME) in forecasts. Instead, it was because they reduced random error or scatter (VE): they made forecasts more consistent.

We anticipated that the cleansed statistical forecasts would improve forecasting more than the non-cleansed ones. However, our rationale for this was based on our expectation that the cleansed forecasts would lower bias by reducing the under-adjustment from the mean of the series – the salient anchor in the unaided and non-cleansed statistical forecast conditions. It was on this basis that we generated hypotheses 2, 3a, and 5. However, no differences in the effectiveness of the cleansed and non-cleansed statistical forecasts were evident, either as main effects or as interactions in our analyses of MAE, ME and VE. They did not affect degree of under-adjustment from the mean of the series.

Provision of cleansed and of non-cleansed statistical hypotheses both improved overall accuracy but there was no difference in the degree to which they did so. This was because there was no difference in the extent to which they reduced VE.

3.3.2. Effects of promotions in the data series and in the periods to be forecast Proportion of promotions in the data series and whether the forecast was for a normal or for a promotional period interacted in their effects on overall forecast accuracy (Table 1): a greater proportion of non-promotional periods in the data specifically helped forecasts for non-promotional periods. To understand why this was, we need to consider the separate analyses of ME and VE.

Forecasters are likely to anchor on the overall mean of the data series (Lawrence & O'Connor, 1992). Fewer promotions meant that that overall mean was closer to the mean value of the non-promotional periods but further from the mean value of the promotional periods. So, with fewer promotions in the data series, a larger adjustment from the overall mean of the series was needed to forecast promotional periods but a smaller adjustment was needed to forecast non-promotional periods. The data show that under-adjustment was greater when a larger adjustment was needed. This is to be expected. In psychophysics, the Weber-Fechner Law (Baird & Noma, 1978; Fechner, 1860; Weber, 1834) summarizes many findings showing that errors in discrimination are proportional to the overall size of the stimulus being judged. Hence, because under-adjustment was proportional to the size of the required adjustment, ME became less positive on normal periods but more negative on promotional ones as the proportion of promotions in the data series decreased (Table 2).

A greater proportion of promotions in the data series increased its variability. If people used their estimate of the overall mean of the series as a judgment anchor, this estimate would have been more variable when the proportion of promotions in the data series was higher. As a result, VE was also higher (Table 2).

To allow for the absence or presence of a promotion in the period to be forecast, people would have had to adjust away from this initial judgment anchor. When there was no promotion planned, this would require forecasters merely to estimate from the data series the mean value of sales when no promotion had occurred (and to move their judgment away from the initial anchor towards that mean value). However, when a promotion was planned, they would have to do more than just estimate the mean value of sales when a promotion occurred: they would also have to take into account the relation between the size of a promotion and the elevating effect it had on sales. This could be done in various ways (e.g., via some kind of mental regression). However, it is reasonable to assume that this additional process would be imperfect and so add some random error to the forecasts. As a result, VE was higher in forecasts for promotional periods (Table 2).

The reasons for the relatively low value of MAE when forecasts for normal (rather than promotional) series were made from series with 10% (rather than 40%) promotions are now clear. VE is reduced by forecasting for normal rather than for promotional periods. Additionally, VE is reduced with fewer promotions in the data series. Finally, fewer promotions in the data also result in a reduction of the size of the positive ME associated with forecasts that are made for normal periods. This combination of two factors reducing VE (normal periods, fewer promotions) and one factor reducing ME (fewer promotions) results in a particularly low MAE value. MAE is higher in all other cases because factors that lower VE and those that lower ME do not combine in the same felicitous manner. For example, consider the case in which forecasts are made for a promotional period from data series containing 40% promotions. Here, the higher proportion of promotions in the data series reduces the size of the negative ME associated with making forecasts for promotional periods. However the beneficial effects of this are counteracted by the fact that VE is higher when forecasts are made for promotional periods and when the proportion of promotions in the data series is higher.

Why did the presence of a statistical forecast lower VE and, hence, MAE? We have argued that forecasters first estimate the overall mean of the data series and that this acts as an initial judgment anchor. Furthermore, this is an error-prone process: VE is higher when the data series is more variable. Both types of statistical forecast act to make it less error-prone. Forecasters could reduce the amount of random error in their estimate of the series mean simply by averaging it with the non-cleansed statistical forecast or by averaging it with the cleansed statistical forecast plus some increment specific to the proportion of promotions in the data series.

3.3.3. Summary We can explain the patterns in the data by assuming that people produce their forecast in two steps. First, they estimate the overall mean of the data series in order to use it as an initial judgment anchor. The size of the random error associated with this estimate is higher when data series are more variable but it can be reduced by provision of a statistical forecast.

Second, forecasters adjust away from this initial anchor to allow for whether a promotion is planned or not. Under-adjustment results in under-forecasting on promotional periods and over-forecasting on normal ones. The size of the under-adjustment is greater when a larger adjustment is required: hence, a greater proportion of promotions in the data series results in greater (positive) ME on normal periods but smaller (negative) ME on promotional periods. Adjustments are based on just the mean value of sales on non-promotional periods when normal periods are forecast but they must take into account the relation between size of promotions and the size of their effects on promotional periods. This additional process is error-prone and hence results in higher VE on promotional periods.

4. Experiment 2: Use of statistical forecasts that take account of effects of promotions

Unexpectedly, Experiment 1 failed to reveal any difference in forecast accuracy between participants who received the cleansed statistical forecasts and those who received the non-cleansed ones. Non-cleansed statistical forecasts are cruder: they require less processing of the data series but always require some adjustment. In contrast, cleansed forecasts provide a clearly defined baseline series and, as a result, they can be accepted without adjustment for non-promotional periods. Despite this, participants made large upward adjustments on these periods (Table 2).

It is possible that people who see that the cleansed statistical forecast does not account for promotions falsely infer that cannot be ‘trusted’ for normal periods either. As a result, they make adjustments for both types of period. Forecasts need to be relevant, justifiable and valuable in dealing with future uncertainties in order for them to be acceptable (Gönül, Önkal, & Lawrence, 2006). The clear unacceptability of the cleansed statistical forecasts on promotional periods may have been inappropriately generalized to affect the acceptability of those forecasts for both types of period (promotional and normal).

This possibility prompted us to carry out Experiment 2. We provided participants with 'optimal' forecasts. Each forecast for a promotional period was based on the cleansed statistical forecasts but with the appropriate increase in sales produced by the promotion in the promotional period added to it. While it is not completely realistic to obtain such forecasts in business practice, it is an approach that is now approximated by recently developed forecasting methods that include promotional modelling (e.g., Huang, et al., 2014; Kourentzes & Petropoulos, 2016; Trapero, et al., 2013).

We suggested above that cleansed forecasts for non-promotional periods are not accepted without adjustment because it is clear to forecasters that cleansed forecasts for promotional periods are unacceptable without adjustment and this leads to a lack of trust in all forecasts. As a result, all forecasts are adjusted. In the present experiment, it was made clear to forecasters that forecasts for promotional as well as for normal periods were acceptable without adjustment. If our suggestion is correct, then forecasters would have no reason not to trust the statistical forecasts. As a result, they should be judged acceptable and adopted without adjustment. More formally,

Hypothesis 6: Optimal forecasts will be accepted without adjustment.

4.1. Method

The experiment was identical to Experiment 1, except that statistical forecasts for promotional periods were elevated by an amount that was appropriate to the size of the planned promotion.

4.1.1. Participants Fifty students from University College participated in the study. Their mean age was 17.77 years (*SD* = 0.87 years) and 40 of them were female.

4.1.2. Stimulus materials, design and procedure In the instructions and in the experiment, the statistical forecasts were presented as shown in Figure 3. In all other respects, the experiment was identical to the first one. The instructions with regard to the statistical forecast were adapted as follows: "*The orange line is a forecast from a statistical model*[2]*. The model is based on the cleaned sales data: the promotion effects have been taken out of the data until only the baseline remained. The model uses these baseline data to produce the statistical forecasts and then adds the promotion effects on top of this forecast. You can see the predictions it made in the past and what it predicts for time period 51 and 52. You can choose whether or not to follow the statistical forecast.*"

Insert Figure 3 about here

4.2. Results

Data for MAE, ME, and VE of forecasts for periods with and without promotions and from data series with low and high frequency of promotions are shown in Table 3. A repeated-measures ANOVA of the MAE revealed a main effect of the frequency of promotions (*F* (1, 49) = 30.15, *p* < .001), indicating that forecasts were more accurate with fewer promotions in the data series, and a main effect of presence of a promotion (*F* (1, 49) = 18.62, *p* < .001), showing that forecasts were more accurate for normal than for promotional periods. Analysis of ME revealed only a main effect of the frequency of promotions (*F* (1, 49) = 26.03, *p* < .001) indicating slight over-forecasting when data series contained 40% promotional periods but slight under-forecasting when they contained 10% promotional periods. Analysis of VE indicated a main effect of the frequency of promotions (F (1, 49) = 17.15, p < .001) and a main effect of the presence of a promotion (F (1, 49) = 13.03, p = .001). The direction of these effects mirrored that of those obtained for MAE.

4.2.1. Comparison of performance with that obtained in Experiment 1 Did the enhanced statistical forecast provided in this experiment lead to better forecasting than that obtained by using unaided judgment or by using judgment aided by the provision of other types of statistical forecast?

To find out, we compared forecast accuracy with that obtained in the three conditions of Experiment 1 (Figure 4). With regard to overall error (MAE), performance was significantly better than unaided judgment (*F* (1, 99) = 27.77, *p* < .001), aided judgment with a non-cleansed-series statistical forecast (*F* (1, 99) = 14.71, *p* < .001) and aided judgment with a cleansed-series forecast (*F* (1, 99) = 23.13, *p* < .001).

To compare the size of ME scores across experiments, we analyzed their absolute values. As hypothesized, those in the present experiment were lower than those in all conditions of the previous experiment: unaided judgment (*t* (58) = -6.12, *p* < .001); judgment aided with non-cleansed statistical forecasts (*t* (63) = -5.46, *p* < .001): judgment aided with cleansed statistical forecasts (*t* (60) = -4.56, *p* < .001)[4]. Similarly, the VE was significantly lower in the current experiment than it was in all conditions of the previous experiment: unaided judgment (*t* (76) = -7, *p* < .001), judgment aided with a non-cleansed statistical forecasts (*t* (68) = -3.02, *p* = .004); judgment aided with cleansed statistical forecasts (*t* (73) = -4.11, *p* < .001)[4].

Insert Table 3 and Figure 4 about here

4.3. Discussion

Provision of optimal statistical forecasts significantly reduced all types of error relative to the corresponding error levels observed in all conditions of Experiment 1. In particular, the absolute size of the directional error reduced very considerably. This implies that under-adjustment from the initial anchor was strongly attenuated. However, MAE scores show that a fair amount of error still persisted (Figure 4). This was primarily driven by VE. Although this type of error was significantly lower than it was in all the conditions of Experiment 1, it remained high at 83% of its size in that experiment. As before, it was larger when there were more promotions in the data series and when promotions were planned for a forecast period. These influences on VE are likely to explain their re-appearance in the analyses of MAE. These results indicate that Hypothesis 6 should be rejected.

Lim and O'Connor (1995, Experiment 3) obtained similar findings to ours. They found that forecasters made insufficient use of near-perfect statistical forecasts that were generated by taking the average of a highly reliable statistical forecast and the actual outcome. Forecasters put too much weight on their own views and not enough on the statistical forecast. Similarly, Gardner and Berry (1995) found that people performing a control task who were freely offered perfectly correct advice decided to obtain it on only 44% of occasions. Furthermore, those who obtained it acted in accordance with it on only 73% of occasions. One interpretation of both of these results is that people tend to be overconfident in their own abilities. As a result, they do not take sufficient account of good advice.

According to the account that we provided of the results from Experiment 1, forecasts are produced in two stages. First, forecasters (even those who are provided with statistical forecasts) make their own assessment of the mean of the data series to use as an initial judgment anchor. This assessment is subject to random error that is reflected in the VE scores. This random error is greater when data series are more variable. They are more variable when they contain a higher proportion of promotions and, hence, VE is greater when the proportion of promotions in the data series is higher. This same effect was found in the present experiment and so it is reasonable to assume that forecasters initially processed the series in a similar way in the present experiment.

The statistical forecasts examined in Experiment 1 were beneficial because they reduced VE. The statistical forecasts used in the present experiment also reduced VE. We suggested that this reduction occurs because forecasters can obtain estimates of the series mean both from the raw data series and from the series of past statistical forecasts. (Unaided judgmental forecasters can use only the data series.) A weighted average of these two estimates then provides the initial judgment anchor. If people are less confident in the statistical forecasts, they may put insufficient weight on the estimate obtained from them. As a result, VE may be reduced but not by as much as it could be. In the present experiment, the reduction in VE was greater than that produced by the statistical forecasts provided in Experiment 1. This may have been because the description of how statistical forecasts were generated provided in the instructions gave forecasters greater confidence in them: as a result, they put more weight on them and thereby generated a more accurate estimate of the series mean to use as an initial judgment anchor.

In the second stage, forecasters adjust away from the initial judgment anchor to take account of the presence or absence of a promotion in the period to be forecast. We saw in Experiment 1 that adjustment is typically insufficient (Tversky & Kahneman, 1974). As a result, promotional periods are under-forecast whereas normal periods are over-forecast. Adjustments for normal periods are based just on the mean value of normal periods in the data series but those for promotional periods have to take account of the relation between the size of promotions and the elevation in sales that they produce. This additional process is error-prone and therefore increases VE of forecasts for promotional periods relative to forecasts for normal periods.

This same effect (higher VE on promotional periods) was found in the present experiment. However, in contrast to Experiment 1, analyses of ME showed that there was no evidence of under-forecasting on promotional periods or of over-forecasting on normal ones. Thus, including an element allowing for promotions in statistical forecasts is beneficial not just because it reduces VE but also because it reduces the absolute size of ME. However, VE was still higher for forecasts for promotional periods than for those for normal ones. This implies that people do not merely accept the statistical forecast. Their low ME scores show that, on average, the mean value of their forecasts for both normal and promotional periods is very close to those provided by the statistical forecasts. However, there is considerable scatter around these mean values and this scatter is greater for forecasts for promotional periods. We attribute this greater scatter to additional error-prone cognitive processing that is needed to allow for the promotion function (i.e., the relation between promotion size and its effect).

Statistical forecasts that include an element to allow for the effects of promotions are beneficial because they reduce both bias and random error in forecasts. However, forecasters do not accept them automatically. In fact, of the 4000 forecasts that participants made in Experiment 2, only 333 (8.33%) were equal to the statistical forecast that they had been given. Thus adjustments were still made and they must have been responsible for the high levels of VE that persisted in this experiment.

Levels of VE were also affected by same variables as they were in Experiment 1: the nature of both the data series (proportion of promotions) and the periods to be forecast (normal or promotional). Because the way that VE levels are affected by these variables when statistical forecasts (of whatever type) are provided is the same as the way in which they are affected in unaided judgmental forecasting, our view is that the provision of statistical forecasts does not fundamentally alter the cognitive processes that forecasters employ to perform their task. Instead, they facilitate these processes and do so more for some of them (e.g., the 'de-biasing' observed in Experiment 2) than for others (e.g., extracting an initial mental anchor from the data series). In other words, forecasters still used an anchoring-and-adjustment heuristic when given optimal statistical forecasts but their estimate of the appropriate anchor is somewhat more consistent and their adjustment from that anchor is almost free of bias.

5. Comparison of participants' performance with that achieved by raw statistical forecasts

Up to this point, we have compared the accuracy of unaided judgmental forecasts with that of judgmental forecasts made after the provision of statistical forecasts of various types. This type of comparison addressed the primary question that motivated the work reported here: is it worth providing statistical forecasts to judgmental forecasters? In this section, we address a different question that is of interest but was not a primary motivator of our work. Does judgmental adjustment of raw statistical forecasts improve forecast accuracy?

To answer this question, we compared the accuracy of judgmental forecasts (made with or without access to statistical forecasts) with the accuracy of raw statistical forecasts. Previous non-experimental research that has analyzed company forecast records has indicated that forecasters tend to make too many adjustments (Frances & Legerstee, 2009) but that, on the whole, these adjustments tend to produce final forecasts that are better than the original raw statistical ones

(Syntetos, Nikolopoulos, Boylan, Fildes & Goodwin, 2009; Syntetos, Nikolopoulos & Boylan, 2010). However, these studies did not examine the effect of the proportion of promotions in the data series.

Insert Table 4 about here

Table 4 shows MAE scores of participants' forecasts and of raw statistical forecasts for series with 40% and with 10% promotions in the four conditions of the two experiments. First consider the three conditions in which statistical forecasts were provided. When those forecasts were based on non-cleansed data, the participants' forecasts were *more* accurate than the raw statistical forecasts when there was a large proportion (40%) of promotions in the data series (t (50) = 4.27; p < .001) but *less* accurate than the raw statistical forecasts when there was a small proportion (10%) of promotions in the data series (t (50) = 7.54; p < .001). The pattern was similar when statistical forecasts were based on cleansed data: participants' forecasts were *more* accurate than the raw statistical forecasts when there was a large proportion (40%) of promotions in the data series (t (50) = 5.71; p < .001) but *less* accurate than the raw statistical forecasts when there was a small proportion (10%) of promotions in the data series (t (50) = 7.21; p < .001). However, in Experiment 2, when statistical forecasts were optimal, participants' forecasts were *less* accurate than the raw statistical forecasts both when there was a large proportion (40%) of promotions in the data series (t (49) = 19.07; p < .001) and when there was a small proportion (10%) of promotions in the data series (t (49) = 17.37; p < .001).

Now consider the condition in Experiment 1 in which no statistical forecasts were provided. As participants in this condition received the same set of data series as those in the other conditions, we can compare their performance with that achieved by the statistical forecasts in the other three conditions. These analyses revealed that, when there was a high proportion (40%) of promotions in the data series, participants' performance in the unaided condition was not significantly different from that achieved by statistical forecasts based on non-cleansed data (t (50) = 1.49; NS) or cleansed data (t (50) = 1.70; NS) but was worse than the performance of optimal forecasts (t (50) = 15.04; p < .001). In contrast, when there was a low proportion (10%) of promotions in the data series, participants' performance in the unaided condition was worse than that achieved by all three types of statistical forecast: those based on non-cleansed data (t (50) = 9.54; p < .001): those based on cleansed data (t (50) = 8.51; p < .001); and the optimal forecasts (t (50) = 12.26; p < .001).

In summary, judgmental adjustment was beneficial only a) when statistical forecasts did not take promotions into account, *and* b) when a high proportion of periods in the data series were affected by promotions. When statistical forecasts did take promotions into account, forecasters made unnecessary adjustments (c.f. Frances & Legerstee, 2009) but, when forecasts did not take promotions into account, their adjustments improved forecasts if there was a high proportion of promotion periods in the data series (c.f. Syntetos et al, 2009, 2010).

6. General discussion

We provided forecasters with different types of statistical forecast to investigate how effective they are in improving forecasters' accuracy. We also varied the type of period (normal versus promotional) to be forecast and the proportion of promotional periods in the data series because we expected these factors to influence the benefits that statistical forecasts bestow on forecasting performance[5]. Finally, we developed an account of how forecasts are made from time series that are perturbed by sporadic events (i.e. promotions) and of how those forecasts are affected when forecasters have access to statistical forecasts. Here we discuss each of these aspects of our work in turn.

6.1. Effects of statistical forecasts on forecast accuracy

Statistical forecasts that take no account of whether periods in the data series were affected by sporadic events, such as promotions, provide the most common form of forecasting support for practitioners (e.g., Fildes, et al., 2009; Trapero, et al., 2013). However, in experimental research (e.g., Goodwin and Fildes, 1999; Goodwin et al, 2011), researchers have investigated the usefulness of statistical forecasts based only on normal periods not subject to promotions. We expected that the latter approach would be more effective in improving forecasting accuracy (Hypothesis 2).

While both of these types of statistical forecast improved accuracy relative to that observed with unaided judgmental forecasting (Hypothesis 1), there was no difference in the degree to which they did so. Given previous work by Lim and O'Connor (1995) and the persuasiveness of the arguments in favour of using statistical forecasts based on cleansed data series, this finding was unexpected. However, the rationale for Hyporthesis 2 was based on the assumption that statistical forecasts reduce bias: we anticipated that the anchoring bias for normal periods would be removed when statistical forecasts are based on cleansed rather than uncleansed series. In fact, our data show that statistical forecasts were effective because they reduced scatter (VE) rather than bias (ME) and there is no reason to expect scatter to be reduced more by statistical forecasts based on cleansed series than by statistical forecasts based on non-cleansed data series.

It appears that statistical forecasts that are clearly inadequate for promotional periods affect the degree to which forecasters feel able to trust them for normal periods (even when they are, in fact, optimal for those periods). We reasoned that statistical forecasts that are optimal for both promotional and normal periods should be seen as trustworthy and therefore be capable of reducing the anchoring biases. Experiment 2 demonstrated that this was so: ME values very close to zero showed that anchoring biases were virtually eliminated. However, VE values remained high at 83% of the level observed in the aided conditions of Experiment 1. Nevertheless, the marked drop in overall error (MAE) levels indicates that efforts to incorporate promotional effects into statistical forecasts (e.g., Huang, et al., 2014; Kourentzes & Petropoulos, 2016; Trapero, et al., 2013) hold great promise for increasing the effectiveness of forecasting support systems.

6.2. Effects of promotions in the periods to be forecast

We expected participants to anchor on the mean level of the data series and to adjust upwards/downwards from this to take account of the presence/absence of a promotion planned for the forecast period. As adjustment is typically insufficient (Tversky & Kahneman, 1974), we expected under-forecasting on promotional periods but over-forecasting on normal ones when forecasting was unaided or supported by a statistical forecast based on non-cleansed data series (Hypothesis 3a). This is indeed what we found, thereby confirming forecasters use of the anchoring heuristic. We expected that this anchoring bias would not be present on normal periods when statistical forecasts were based on cleansed data series as forecasters would realise that the statistical forecast could be accepted without adjustment (Hypothesis 3b). However, as we discussed in the previous section, these forecasts appear not to have been trusted (perhaps because those for promotional periods obviously needed adjustment). Forecasters continued to use the mean of the series as a judgment anchor and adjust down from it (insufficiently) to make forecasts for normal periods. Hence, over-forecasting for those periods persisted.

6.3. Effects of proportion of promotions in the data series

We expected that a lower proportion of promotional periods in the data series would reduce overall forecasting error on normal periods but increase it on promotional ones (Hypothesis 4). In fact, lowering the proportion of promotions resulted in a lower MAE on normal periods but promotional ones were unaffected. Decomposing overall error showed why this was so. On promotional periods, the absolute size of the under-forecasting bias increased when the proportion of promotions in the data series was reduced but scatter decreased. These two effects cancelled one another out and so there was no resultant effect on overall error. (For normal periods, reducing the proportion of promotions in the data series decreased both the over-forecasting bias and scatter: hence, the predicted effect occurred.)

When there were fewer promotional periods in the data series, statistical forecasts derived from non-cleansed series were closer to the baseline forecasts provided by the statistical forecasts derived from cleansed data series. Hence we expected any accuracy advantage of the statistical forecasts based on cleansed series (over the statistical forecasts based on non-cleansed series) would be greater when the proportion of promotions in the data series was higher (Hypothesis 5). However, there was no evidence of an interaction between proportion of promotions in the data series and type of statistical forecast. As we have seen, forecasters in Experiment 1 appear to have made their judgments in a similar way whether they were unaided or supported by either type of statistical forecast. The only reason that statistical forecasts helped was that they enabled them to make these judgments more consistently.

6.4. Forecasting from time series subject to sporadic perturbation

We have suggested that the cognitive processes underlying forecasting from time series subject to sporadic perturbation are broadly the same whether or not forecasting is aided by provision of statistical forecasts. This is particularly true for the two types of statistical forecast in current use: those that take no account of whether periods in the data series are normal or promotional and those that are based only on the normal periods. As Experiment 1 showed, anchoring effects and effects of proportion of promotions in the data series were unaffected by the presence of a statistical forecast or by its type when present. This implies that the way that the judgments were made was the same across all conditions of Experiment 1. The provision of statistical forecasts did improve accuracy but this was because they made judgment processes more consistent rather than because they changed the nature of those processes.

The optimal statistical forecasts provided in Experiment 2 virtually eliminated under-adjustment. However, VE values remained high. Furthermore, they were still affected by variables that affected VE in Experiment 1. We suspect that similar cognitive processes were responsible for performance in the two experiments. A mental anchor based on the mean of the data series was first extracted. This process was based on noisier data when the series contained more promotions, thereby explaining the effect of that variable on VE. The optimal forecasts ensured that, on average, the adjustments from this anchor were appropriate. However, VE was still higher when forecasts had to be made for promotional periods. To us, this implies that the adjustment process was more complex on promotional periods than on normal ones (because of the additional processing stage involved in extracting and using the relation between the size of a promotion and the effect that it had). Clearly, optimal statistical forecasts are not accepted automatically. They influence judgment but do not supersede it.

Why was there a considerable level of variable error, regardless of the presence and type of statistical forecast? Human forecasters introduce inconsistency or random error into forecasts. At least in part, this error is likely to arise from the noise that is inherent in cognitive processing. Since Thurstone (1926), it has been known that judgment contains a random element. When people make a series of judgments about a criterion variable (e.g., salary levels of a number of different people) from information they are given about cue variables imperfectly correlated with the criterion (e.g., the weight, age, and nationality of those people), the relation between their judgment and the cues contains a random element (Brehmer, 1978) that decreases but does not disappear with practice and feedback. There are many hypotheses about why this occurs (Harvey, 1995). For example, Hammond & Summers (1972) referred to a failure of cognitive control: just as hand tremor causes inconsistency in the execution of fine motor skills, so some analogous process is affects judgment. Modern computational modelling of cognition is based on the notion that each component process contributes some random error to the total observed in the data (Lewandowsky & Farrell, 2011).

Noise inherent in cognitive processing is unlikely to be the only reason for high VE levels. Lawrence et al (2006, p 501) suggest that small damaging adjustments of the sort reported by Fildes et al (2009) may reflect “a tendency to tinker at the edges”. In other words, forecasters *intend* to introduce these small changes that do not, overall, lead to (greater) over-forecasting or under-forecasting but do increase scatter. But why would forecasters do this?

There are various possibilities. One is that the changes that they make provide them with a way in which to assert their ‘ownership’ of the forecasts (Önkal & Gönul, 2005). Another concerns people’s responses to automation. Whenever tasks become partially automated, concerns tend to arise among those responsible for performing them that they risk becoming de-skilled (Bainbridge, 1983). Without feedback about the effects of their own actions, they will not be able to acquire or maintain the abilities that they need to perform their tasks autonomously (something that may be needed if the automated system suddenly becomes unavailable). Hence, to ensure they receive such feedback, operators may occasionally over-rule or interfere with the output produced by the automatic system. (For forecasters, receiving feedback about statistical forecasts is no substitute for receiving it about the forecasts that they have generated themselves: only in the latter case is the rationale for the forecasts known.)

6.5. Practical implications

A number of practical implications follow from these results. First, our main message for practitioners is that the provision of statistical forecasts reduces forecast error but whether those statistical forecasts are based on data cleansed of promotional effects does not matter. This knowledge could save time and money because it implies that cleansing the data, a process that is itself subject to biases (Webby, O'Connor, & Edmundson, 2005), is unnecessary. Even a relatively simple statistical forecast can be of value for a company. Hence, companies that wish to improve their forecasting accuracy but do not currently have a large budget or manpower to spare can still benefit from a simple approach that requires minimal effort.

The second experiment indicates that forecasting accuracy can greatly benefit from statistical forecasts that incorporate promotion effects. Importantly, this means that practitioners should come to grips with the new developments in forecasting research. Early adoptors can have a significant competitive advantage resulting from an improved forecasting accuracy. Additionally, this has also implications for forecast support system developers. The incorporation of regression based models of promotions (e.g., Huang, et al., 2014; Kourentzes & Petropoulos, 2016; Trapero, et al., 2013) should be integrated in future statistical forecasting software.

6.6. Limitations and paths for future research

Our study is subject to limitations. Some of these suggest avenues for future research.

6.6.1. More complex series and other approaches to forecasting We used relatively simple series and statistically forecast them using an exponential smoothing approach. It is possible that with more complex (e.g., seasonal) series and other statistical approaches to forecasting (e.g., ARIMA), results might have been different. While we agree that more appropriate forecasting methods can reduce bias in the final forecast (as Experiment 2 demonstrated), we have seen that they have little effect on variable error in the final forecast (again, as Experiment 2 showed). Hence, we expect that our overall conclusions will be generalizable to a range of combinations of different series and forecasting methods that vary in appropriateness.

Our use of exponential smoothing should ensure that our findings are relevant to practitioners. Many surveys have shown that it is easily the most dominant approach used by business practitioners. Mentzer and Kahn's (1995) survey of 478 organizations revealed that exponential smoothing was used by 92% of them. In Sanders and Manrodt's (1994) survey of 96 companies, it was the second most commonly used quantitative technique after a simple moving average. As Goodwin (2010, p 33), pointed out: "Fifty years on, researchers are still finding ways to improve the Holt-Winters method and to extend the conditions where it can be applied. This continued interest is a testament to the method's ability to produce reliable forecasts without sacrificing simplicity or transparency".

6.6.2. Promotions with other characteristics In the future, it would be useful to examine forecasts for promotional periods subject to promotion functions with other properties: noise, post-promotion effects, and non-linearity of promotion functions. First, the promotion function we used was noise-free. In practice, promotion functions are likely to be noisy. Noisy promotion functions are likely to increase the complexity of the cognitive processes needed to identify them and so increase the variable error associated with those processes. This would be likely to further impair forecast accuracy on promotional periods relative to that on normal ones. Second, in our experiments, promotions had effects only on the periods on which they were applied. In practice, effects of promotions may extend beyond that. For example, by bringing forward people's one-off purchases to the promotion period, they would increase sales for that period but decrease them for the following one(s). Again, this would likely to increase the complexity of the cognitive processes needed to identify the effects of promotions, make those processes more error prone, and thereby add to the variable error of forecasts for promotional periods. Finally, in practice, promotion functions may be non-linear (e.g., sigmoid in shape) whereas they were linear in experiments reported here. Given that human judges tend to linearize non-linear functions (e.g., Brehmer & Slovic, 1980), sigmoid promotion functions are likely to be associated with over-estimation of the effects of large promotions. This would make mean error less negative in those particular cases.

We examined forecasting from data series containing 10% and 40% promotional periods. We saw that a lower frequency of promotions produced less overall error on non-promotional periods but not on promotional ones. This was because largely because over-forecasting was lower on non-promotional periods when there were fewer promotions. We attributed this to the fact that little adjustment from the anchor (i.e., the mean of the series) was needed for those periods. So what would we expect if we were to examine forecasting from data series containing, say, 90% promotional periods? Would the effect be reversed because now little adjustment would be needed for promotional periods but a lot would be needed for non-promotional ones? One could certainly argue that, compared with Experiment 1, ME should be larger for non-promotional periods but smaller for promotional ones because more adjustment would now be needed for the former but less for the latter. However, such a high proportion of promotions would produce a much more variable series because the size of promotions itself varies. Extracting the mean of the series to act as the judgment anchor would be more difficult: as a result, we could expect an increase in VE for both non-promotional and promotional periods. This could lead to MAE being higher than in Experiment 1 for both types of period.

6.6.3. Other types of forecaster We used student participants. It could be argued that experts have more insight into how statistical forecasts should be used. However, previous work has shown that experts are subject to similar errors in reasoning as those that afflict novices. Indeed, in some cases, research has even revealed inverse expertise effects (Önkal & Muradoğlu, 1994; Yates, McDaniel, & Brown, 1991). Advice discounting (ignoring or under-weighing the 'advice' of the statistical forecast) may be even greater in experts because they value their own opinion even more than novices do. In fact, Önkal and Muradoğlu (1994) demonstrated that experts exhibited even more over-confidence in their forecasts than those who were less expert. This situation is typical of what happens when experience at a task (e.g., forecasting) fails to produce learning as quickly as people expect it to (Harvey & Fischer, 2005).

An experiment is a simulation or model of a task performed by practitioners. As with any model, some features of the real world task are excluded. Thus we do not expect to see all characteristics of practitioner performance reflected in experimental results. Analysis of data obtained from organizations has revealed that forecasters are often subject to optimism effects: inappropriate upward adjustments of statistical forecasts are greater or made more often than inappropriate downward ones (e.g., Fildes et al, 2009). We did not observe such optimism in our experiments. They were not designed to study or reveal it. All the same, it is possible to argue that optimism would have produced less under-forecasting on promotional periods than over-forecasting on normal periods. We did not find this pattern in the data. However, this prediction does not compare like with like. As we have emphasized, processes underlying forecasting on promotional periods are different from those that underlie it on normal ones. To research into optimism experimentally, further studies should be specifically designed with that aim in mind. One approach is to compare two groups performing exactly the same forecasting task but to label the variable being forecast as ‘profits’ in one case but ‘losses’ in the other. Forecasts are systematically higher in the former case (Harvey & Reimers, 2013).

6.6.4. Increasing the acceptability of statistical forecasts An additional avenue for further research is indicated by the results of the second experiment, which demonstrated that highly sophisticated statistical forecasts that explicitly take account of the effects of promotions benefit forecasters considerably more than those that do not. Further research efforts designed to develop ways of producing such forecasts (e.g., Huang, et al., 2014; Kourentzes & Petropoulos, 2016; Trapero, et al., 2013) are clearly worthwhile. However, this second experiment also showed that even forecasters who are given optimal statistical forecasts make adjustments that impair accuracy.

As we have seen, there are different ways of explaining this finding but they all imply that, for one reason or another, forecasters are not good at taking ‘advice’ from a statistical model. Such discounting of advice has been reported before (e.g., Goodwin, 2000; Lim & O'Connor, 1995) and factors that have been proposed to account for it include concerns about the credibility of a statistical model rather than a human being as a source of advice (Önkal, Gönul, & Lawrence, 2008; Önkal, et al., 2009), and people’s beliefs that their own opinions are better founded than those of others (Harvey & Harries, 2004).

Preventing damaging adjustments has been an important topic in judgmental forecasting. Goodwin et al. (2011) found that neither restriction nor guidance improved accuracy. Indeed, guidance was met with resistance by forecasters. Such resistance is consistent with Bainbridge’s (1983) views about responses to automation. However, as we pointed out, reasons that forecasters make damaging adjustments may not be purely volitional (i.e., arising because, for one reason or another, they *want* to make those adjustments) but may also be at least partly cognitive (i.e., noise may be inherent in the cognitive processes that underlie forecasting). Further research into the facilitation of accepting statistical forecasts is needed.

6.7. Conclusions

Provision of statistical forecasts, even crude ones, can improve forecasting accuracy by reducing variable error. When forecasts are made from time series perturbed by sporadic exogenous events, the effort needed to produce forecasts cleansed of their effects appears not be warranted. However, current efforts to develop methods to incorporate effects of these events into statistical forecasts are worthwhile and are likely to result in improved forecast accuracy.

Footnotes

[1] Nevertheless, we will report comparisons between judgmentally adjusted forecasts and raw statistical forecasts in section 5.

[2] In this paper, colours have been converted to greyscale. Consequently, in the examples of screen displays that we provide, the sales series is shown as the dark grey line, the statistical forecast as a light grey line, and the promotional expenditure as a light grey bar.

[3] Promotional increases and sales were on the same scale in all graphs that participants saw. In such circumstances, it is appropriate to use scale-dependent measures because their meaning is immediately transparent (Hyndman & Koehler, 2006). Scale-independent measures often suffer from asymmetry and they are not able to deal well with values close to zero (Hyndman and Koehler, 2006). Given these considerations, we opted to use scale-dependent error measures.

[4] In cross-experimental comparisons of ME and VE, Levene's test indicated unequal variances and so degrees of freedom were adjusted accordingly.

[5] We also varied type of series (independent versus autoregressive) and forecast horizon (one-step-ahead versus two-steps-ahead). These variables were not germane to our hypotheses, were included only to increase the generality of our conclusions, and were not part of our main analyses. However, across all conditions and both experiments, MAE increased with forecast horizon ($F(1, 204) = 104.57$; $p < .001$) as expected on the basis of decreased predictability and error accumulation (Harvey, 1995; Theocharis & Harvey, 2016). It was also higher for independent than for the autoregressive series ($F(1, 204) = 27.10$; $p < .001$), as expected on the basis of previous work (Reimers & Harvey, 2011).

References

Alvarado-Valencia, J., & Barrero, L. H. (2014). Reliance, trust and heuristics in judgmental forecasting. *Computers in Human Behavior, 36*, 102 - 113.

Arkes, H. R. (2001). Overconfidence in judgmental forecasting. In J. S. Armstrong (Ed.), *Principles of forecasting*. Boston: Kluwer Academic Publishers.

Armstrong, J. S. (2001). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting: a* handbook for researchers and practitioners (pp. 417 - 439). New York: Kluwer.

Bainbridge, L. (1983). Ironies of automation. *Automatica,* 19, 775 - 779.

Baird, J. C., & Noma, E. (1978). *Fundamentals of scaling and psychophysics*. New York: Wiley.

Blattberg, R. C., & Hoch, S. J. (1990). Database models and managerial intuition: 50% model + 50% manager. *Management Science, 36*, 887 - 899.

Bovi, M. (2009). Economic versus psychological forecasting. Evidence from consumer confidence surveys. *Journal of Economic Psychology, 30*, 563 - 574.

Brehmer, B. (1978). Response consistenty in probabilistic inference tasks. *Organizational Behavior and Human Decision Processes, 22*, 103 - 115.

Brehmer, B. & Slovic, P. (1980). Information integration in multiple-cue judgments. *Journal of Experimental Psychology: Human Perception and Performance, 6,* 302 - 308.

Durand, R. (2003). Predicting a firm's forecasting ability: the roles of organizational illusion of control and organizational attention. *Strategic Management Journal, 24*, 821 - 838.

Fechner, G. T. (1860). *Elemente der psychophysik [Elements of psychophysics] (Volume 1).* Leipzig: Breitkopf und Harterl.

Fildes, R., & Goodwin, P. (2007). Against your better judgment? How organizations can improve their use of management judgment in forecasting. *Interfaces, 37*, 570 - 576.

Fildes, R., & Goodwin, P. (2013). Forecasting support systems: What we know, what we need to know. *International Journal of Forecasting, 29*, 290 - 294.

Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting, 25*, 3 - 23.

Fildes, R., & Petropoulos, F. (2015). Improving forecast quality in practice. *Foresight: the International Journal of Applied Forecasting, 36*, 5 - 12.

Frances, P.F. & Legerstee, R. (2009). Properties of expert adjustments on model-based SKU-level forecasts. *International Journal of Forecasting, 25,* 35 - 47.

Gardner, D. H., & Berry, D. C. (1995). The effect of different forms of advice on the control of a simulated complex system. *Applied Cognitive Psychology*, *9,* 555 - 579.

Gönül, M. S., Önkal, D., & Lawrence, M. (2006). The effects of structural characteristics of explanations on use of a DSS. *Decision Support Systems, 42*, 1481 - 1493.

Goodwin, P. (2000). Improving the voluntary integration of statistical forecasts and judgment. *International Journal of Forecasting, 16*, 85 - 99.

Goodwin, P. (2010). The Holt-Winters approach to exponential smoothing: 50 years old and still going strong. *Foresight, Issue 19*, 30 - 34.

Goodwin, P., & Fildes, R. (1999). Judgmental forecasts of time series affected by special events: Does providing a statistical forecast improve accuracy? *Journal of Behavioral Decision Making, 12*, 37 - 23.

Goodwin, P., Fildes, R., Lawrence, M., & Nikolopoulos, K. (2007). The process of using a forecasting support system. *International Journal of Forecasting, 23*, 391 - 404.

Goodwin, P., Fildes, R., Lawrence, M., & Stephens, G. (2011). Restrictiveness and guidance in support systems. *Omega : The International Journal of Management Science, 39*, 242 - 253.

Hammond, K. R. & Summers, D. A. (1972). Cognitive control. *Psychological Review, 79,* 58 - 67.

Harvey, N. (1995). Why are judgments less consistent in less predictable task situations? *Organizational Behavior & Human Decision Processes, 63*, 247 - 263.

Harvey, N. and Fischer, I. (2005). Development of experience-based judgment and decision-making: The role of outcome feedback. In T. Betsch and S. Haberstroh (Eds). *The Routines of Decision-Making.* Erlbaum: Mahwah, NJ, pp. 119–137.

Harvey, N., & Harries, C. (2004). Effects of judges' forecasting on their later combination of forecasts for the same outcomes. *International Journal of Forecasting, 20*, 391 - 409.

Harvey, N., & Reimers, S. (2013). Trend damping: under-adjustment, experimental artifact, or adaptation to features of the natural environment? *Journal of Experimental Psychology, 39*, 589 - 607.

Hilary, G., & Hsu, C. (2011). Endogenous overconfidence in managerial forecasts. *Journal of Accounting and Economics, 51*, 300 - 313.

Huang, T., Fildes, R., & Soopramanien, D. (2014). The value of competitive information in forecasting FMCG retail product sales and the variable selection problem. *European Journal of Operational Research, 237*, 738 - 748.

Hyndman, R. J. & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22,* 679 - 688.

Kotteman, J. E., Davis, F. D., & Remus, W. (1994). Computer-assisted decision making: performance, beliefs, and the illusion of control. *Organizational Behavior & Human Decision Processes, 57*, 26 - 37.

Kourentzes, N., & Petropoulos, F. (2016). Forecasting with multivariate temporal aggregation: The case of promotional modelling. *International Journal of Production Economics, 181*, 145 - 153.

Lawrence, M. (2000). Editorial: What does it take to achieve adoption in sales forecasting? *International Journal of Forecasting, 16,* 147 - 148.

Lawrence, M. and O’Connor, M. (1992). Exploring judgmental forecasting. *International Journal of Forecasting, 8 ,* 15 - 26.

Lawrence, M., Goodwin, P., O’Connor, M., & Önkal, D. (2006). Judgmental forecasting: A review of progress over the last 25years. *International Journal of Forecasting, 22,* 493 - 518.

Lewandowsky, S. & Farrell, S. (2011). *Computational modeling in cognition: Principles and practice.* London: Sage.

Libby, R., & Rennekamp, K. (2012). Self-serving attribution bias, overconfidence, and the issuance of management forecasts. *Journal of Accounting Research, 50*, 197 - 231.

Lim, J. S., & O'Connor, M. (1995). Judgmental adjustment of initial forecasts: Its effectiveness and biases. *Journal of Behavioral Decision Making, 8*, 149 - 168.

Lim, J. S., & O'Connor, M. (1996). Judgmental forecasting with time series and causal information. *International Journal of Forecasting, 12*, 139 - 153.

Mentzer J.T. & Kahn, K. (1995).Forecasting technique familiarity, satisfaction, usage and application. *Journal of Forecasting, 14*, 465 - 476.

Önkal, D., & Gönul, M. S. (2005). Judgmental adjustment: A challenge for providers and users of forecasts. *Foresight: The International Journal of Applied Forecasting, 1*, 13 - 17.

Önkal, D., Gönul, S., & Lawrence, M. (2008). Judgmental adjustments of previously adjusted forecasts. *Decision Sciences, 39*, 213 - 238.

Önkal, D., Goodwin, P., Thomson, M., Gönul, S., & Pollock, A. (2009). The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making, 22*, 390 - 409.

Önkal, D., & Muradoğlu. (1994). Evaluating probabilistic forecasts of stock prices in a developing stock market. *European Journal of Operational Research, 74*, 350 - 358.

Önkal, D., Sayim, K. Z., & Lawrence, M. (2012). Wisdom of group forecasts: Does role-playing play a role? *Omega, 40*, 693 - 702.

Reimers, S. and Harvey, N. (2011). Sensitivity to autocorrelation in judgmental time series forecasting. *International Journal of Forecasting*, 27, 1196 - 1214.

Sanders, N. R. & Manrodt, K. B. (1994). Forecasting practices in US corporations: Survey results. *Interfaces, 24,* 92 - 100.

Sanders, N. R., & Manrodt, K. B. (2003). The efficacy of using judgmental versus quantitative forecasting methods in practice. *Omega, 31*, 511 - 522.

Syntetos, A. A. , Nikolopoulos, K, Boylan, J. E., Fildes, R. & Goodwin, P. (2009). The effects of integrating management judgment into intermittent demand forecasts. *International Journal of Production Economics, 118*, 72 - 81.

Syntetos, A. A. , Nikolopoulos, K, & Boylan, J. E. (2010). Judging the judges through accuracy-implication metrics: The case of inventory forecasting. *International Journal of Forecasting, 26*, 134 - 143.

Theocharis, Z. and Harvey, N. (2016). Order effects in judgmental forecasting. *International Journal of Forecasting, 32*, 44 - 60.

Thurstone, L. L. (1926). The scoring of individual performance. *Journal of Educational Psychology, 17*, 446 - 457.

Trapero, J. R., Pedregal, D. J., Fildes, R., & Kourentzes, N. (2013). Analysis of judgmental adjustments in the presence of promotions. *International Journal of Forecasting, 29*, 234 - 243.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124 - 1131.

Webby, R., O’Connor, M., & Edmundson, R. (2005). Forecasting support systems for the incorporation of event information: An empirical investigation. *International Journal of Forecasting, 21*, 411 - 423.

Weber, E. H. (1834). *De pulsu, resorptione, auditu et tactu* [On stimulation, response, hearing and touch]. Annotationes, anatomical et physiological. Leipzig: Koehler.

Yates, J. F., McDaniel, L. S., & Brown, E. S. (1991). Probabilistic forecasts of stock prices and earnings: The hazards of nascent expertise. *Organizational Behavior & Human Decision Processes, 40*, 60 - 79.

Table 1

Experiment 1: Mean values of participants' mean absolute error (MAE) for each level of promotion frequency and for each type of promotion period in the three conditions of the experiment.

Independent Variables		Statistical forecast			
		None	Cleansed	Not cleansed	Mean
40% promotions	Promotion	33.48	30.13	30.24	31.28
	No promotion	35.06	28.66	31.67	31.79
10% promotions	Promotion	34.54	31.11	29.43	31.69
	No promotion	29.94	24.44	23.73	26.03

Table 2

Experiment 1: Mean values of mean error (ME) and variable error (VE) for each level of promotion frequency and for each type of promotion period in the three conditions of the experiment.

Independent Variables		ME				VE			
		No SF	Cleansed	Not cleansed	Means	No SF	Cleansed	Not cleansed	Means
40% promotions	Promotion	-8.07	-6.47	-3.85	-6.13	32.75	30.41	29.50	30.89
	No promotion	18.21	14.08	18.38	16.89	29.92	28.01	28.51	28.81
10% promotions	Promotion	-10.93	-11.39	-7.04	-9.79	32.29	29.35	28.50	30.04
	No promotion	12.63	8.07	9.84	10.18	29.14	25.58	23.40	26.04

Table 3

Experiment 2: Mean values of mean absolute error (MAE), mean error (ME) and variable error (VE) for each level of promotion frequency and for each type of promotion period.

Independent Variables		MAE	ME	VE
40% promotions	Promotion	25.61	1.99	26.43
	No promotion	23.09	2.54	24.22
10% promotions	Promotion	23.02	-2.8	23.98
	No promotion	19.50	-.59	21.09

Table 4

Mean absolute error (MAE) scores of participants' forecasts and of raw statistical forecasts for series with 40% and with 10% promotions in the four conditions of the two experiments.

Independent Variables		Statistical forecast condition			
		None	Not Cleansed	Cleansed	Optimal
40% promotions	Participants' forecast	34.43	31.10	29.25	24.10
	Statistical forecast	N.A.	36.56	36.86	12.87
10% promotions	Participants' forecast	30.40	24.30	25.11	19.85
	Statistical forecast	N.A.	13.66	15.47	8.88

Figure captions

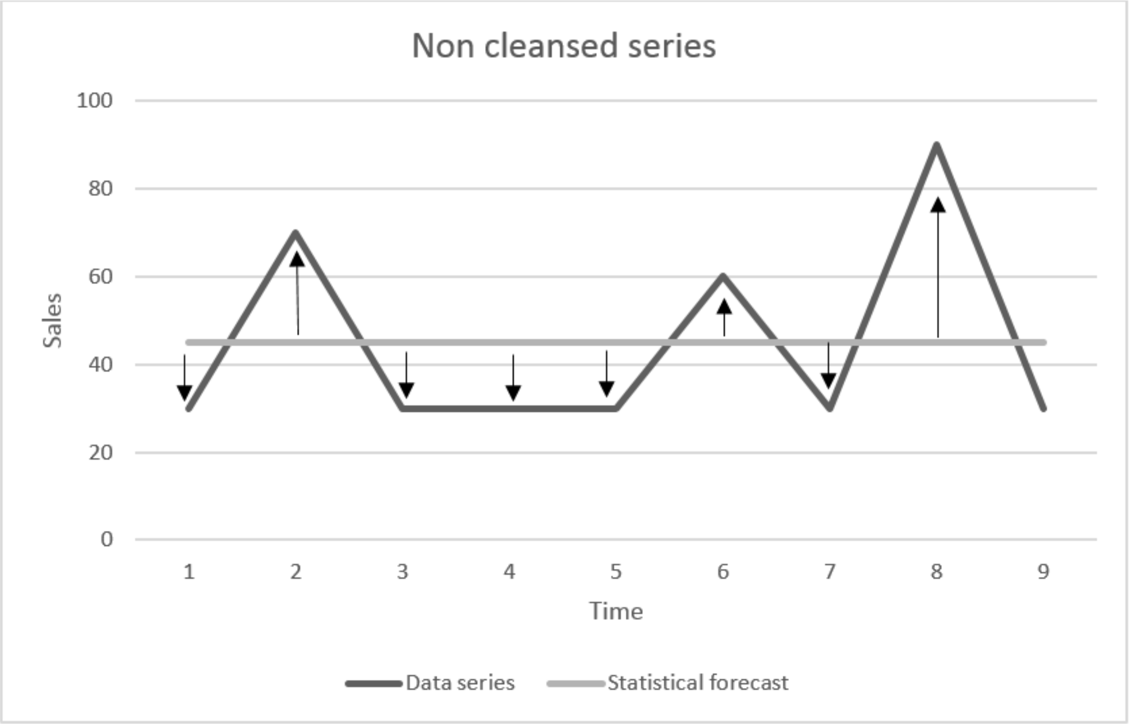
Figure 1. Adjustments necessary for a statistical forecast based on non-cleansed series (upper panel) and cleansed series (lower panel).

Figure 2. Experiment 1: a) Example of screen displays in the unaided group; b) Example of screen displays in the group aided by statistical forecasts based on non-cleansed data series; c) Example of screen displays in the group aided by statistical forecasts based on cleansed data series.

Figure 3. Experiment 2: Example of the screen display.

Figure 4. Error scores associated with different forecasting conditions studied in the two experiments: MAE (upper panel); ME (central panel); VE (lower panel)

Figure 1


Non cleansed series
100
80
60
40
20
0
Sales
1
2
3
4
5
6
7
8
9
Time
Data series
Statistical forecast

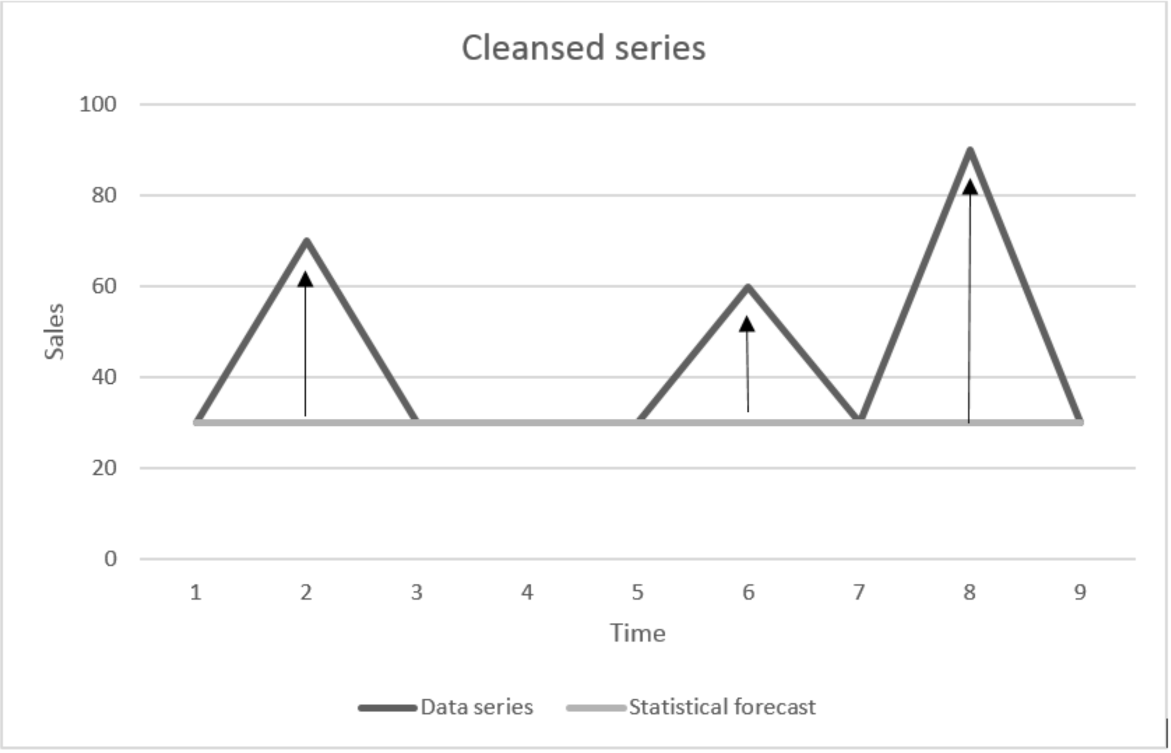
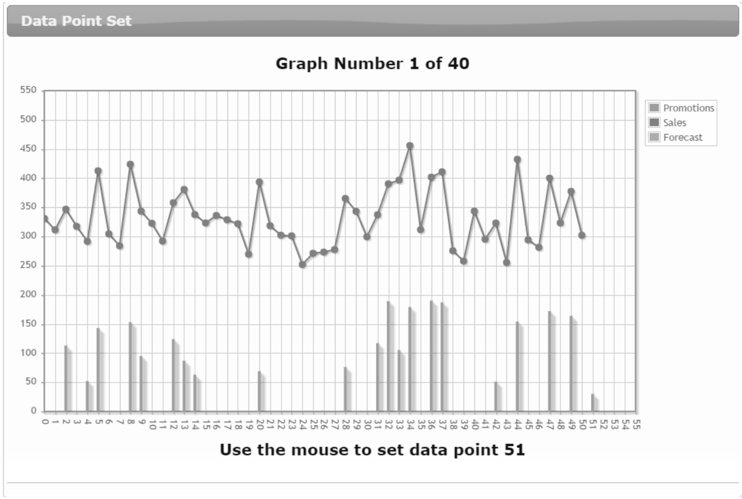

Cleansed series
100
80
60
40
20
0
Sales
1
2
3
4
5
6
7
8
9
Time
Data series
Statistical forecast

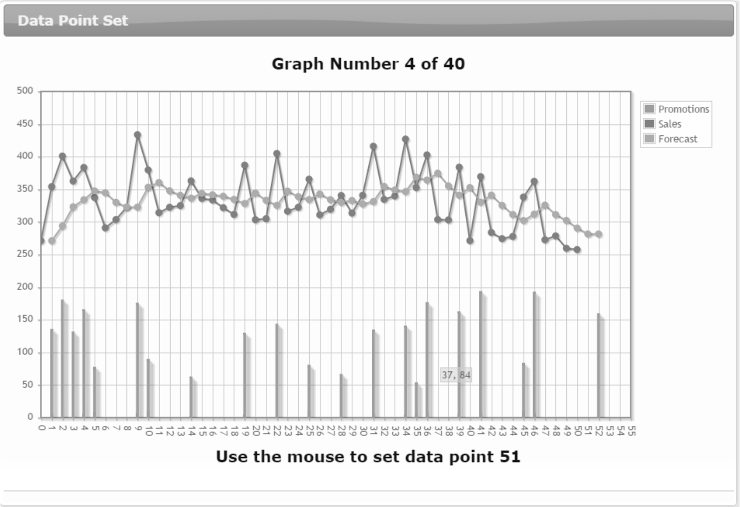
Figure 2a



Graph Number 1 of 40

Use the mouse to set data point 51

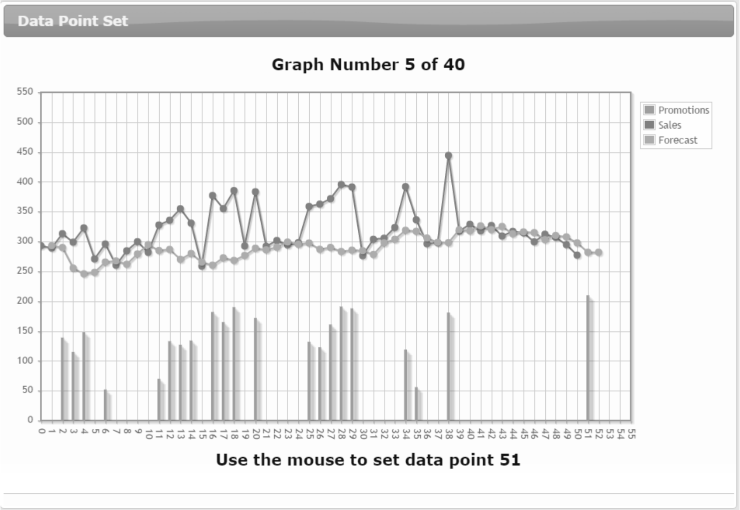
Figure 2b



Graph Number 4 of 40

Use the mouse to set data point 51

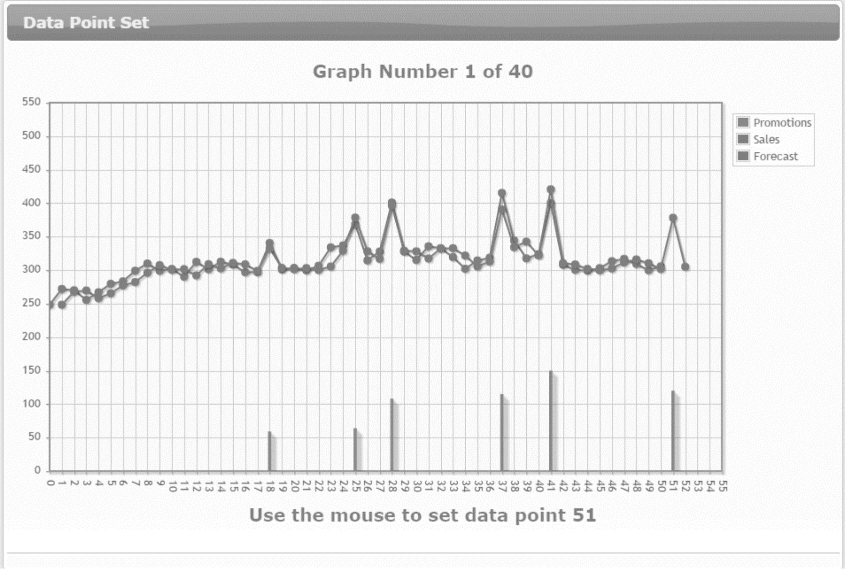
Figure 2c



Graph Number 5 of 40

Use the mouse to set data point 51

Figure 3

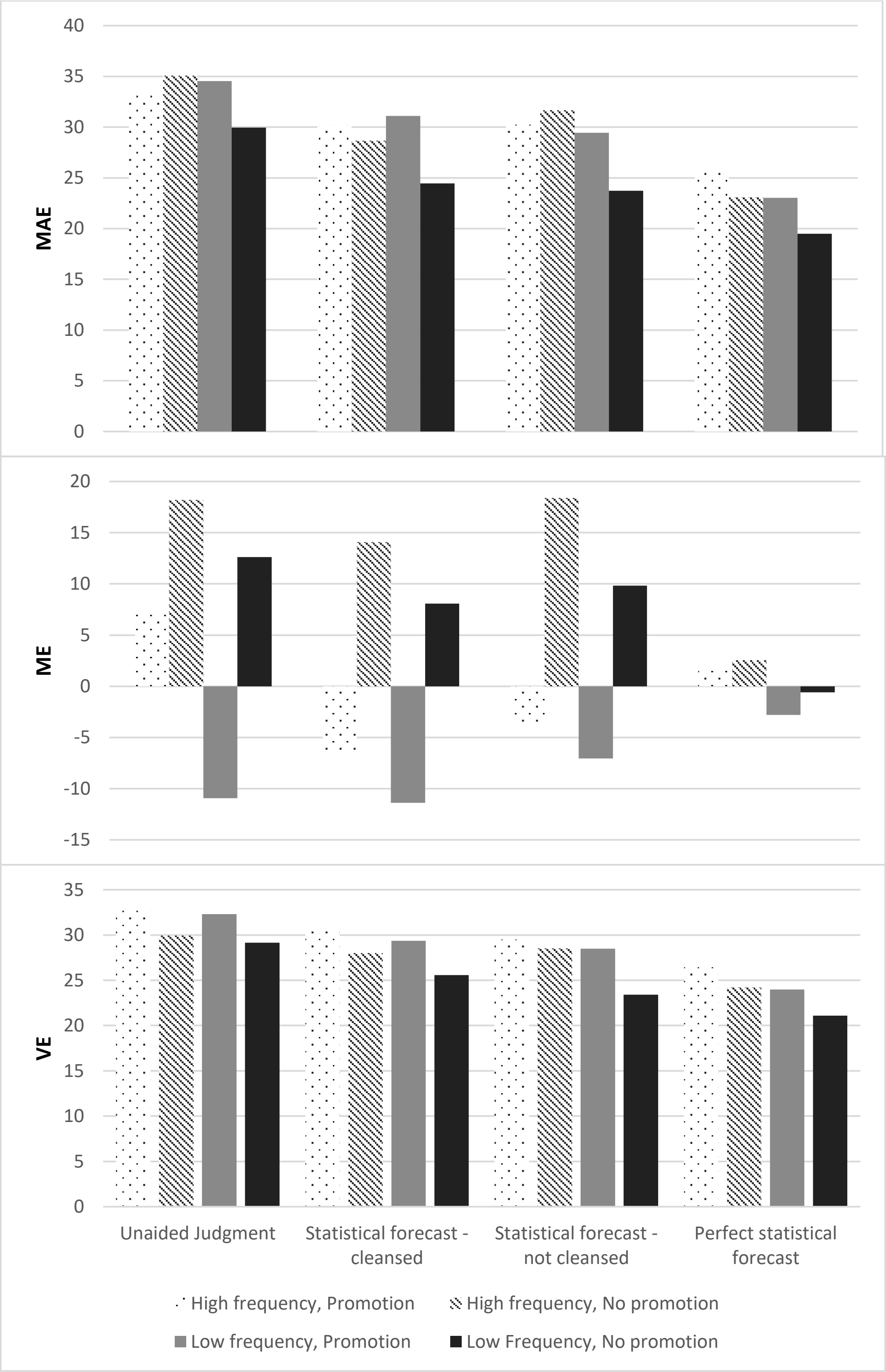


Data Point Set

Graph Number 1 of 40

Use the mouse to set data point 51

Figure 4


MAE
40
35
30
25
20
15
10
5
0
ME
20
15
10
5
0
-5
-10
-15
VE
35
30
25
20
15
10
5
0
Unaided Judgment
Statistical forecast - cleansed
Statistical forecast - not cleansed
Perfect statistical forecast
High frequency, Promotion
High frequency, No promotion
Low frequency, Promotion
Low Frequency, No promotion

APPENDIX – Instruction sheets for Experiment 1

Group A (unaided judgment, no statistical forecast) were given the following text on an instruction sheet: "*Please read this document carefully before you start with the first graph! In this experiment, you will receive a number of graphs such as the one depicted below. On the X-axis, you will find the time period, ranging from 0 to 55. On the Y-axis, you will find the sales number, ranging from 0 to 500. The grey line indicates the sales data of a product in the past 50 time periods. The blue bars*[2] *indicate the promotional investment (e.g., an advertisement campaign) made for that product. The number of promotions can vary: some graphs will have 5 promotions, others will have 20. It is your job to predict the sales number of the following two time periods (51 and 52), as accurately as possible. Pay attention, because sometimes there is a promotion present and sometimes there isn't. You can make your prediction by clicking with your mouse on the graph. An information box with your mouse's location appears next to your cursor. First click on your prediction for time period 51 and only then for time period 52. Afterwards, a box 'next graph' will appear on the bottom of the page."*

Participants in group B (statistical forecast based on cleansed series) saw the following additional text: *"The orange line*[2] *is a forecast from a statistical model. The model is based on the cleaned sales data: the promotion effects have been taken out of the data until only the baseline remained. The model uses these baseline data to produce the statistical forecasts. You can see the predictions it made in the past and what it predicts for time period 51 and 52. You can choose whether or not to follow the statistical forecast"*.

For those in group C (statistical forecast based on non-cleansed data), the additional text was as follows: "*The orange line is a forecast from a statistical model. We have fed the sales data to a statistical model. You can see the predictions it made in the past and what it predicts for time period 51 and 52. This statistical model is a simple model that ignores whether or not a promotion* *took place: it is just based on the value of the sales figures. You can choose whether or not to follow the statistical forecast.*”